Included in the book: *Meeting on Fundamental Physics 'Alberto Galindo'*,
Álvarez-Estrada R. F. *et al.* (Ed.),
Madrid: Aula Documental, 2004.

Do theoretical physicists care about the protein-folding problem?

José L. Alonso[1,2], Gregory A. Chass[1,2,3], Imre G. Csizmadia[3,4,5], Pablo Echenique[1,2] † and Alfonso Tarancón[1,2]

[1] *Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, Pedro Cerbuna 12, 50009, Zaragoza, Spain.*
[2] *Instituto de Biocomputación y Física de los Sistemas Complejos (BIFI), Edificio Cervantes, Corona de Aragón 42, 50009, Zaragoza, Spain.*
[3] *Global Institute Of COmputational Molecular and Material Science (GIOCOMMS), 158-2 Major St., Toronto, ON, M5S 2L2 Canada.*
[4] *Department of Chemistry, University of Toronto, 80 St. George St., Toronto, ON, M5S 3H6 Canada.*
[5] *Department of Medical Chemistry, University of Szeged, 8 Dóm tér, Szeged, H-6720, Hungary.*

Abstract.

The prediction of the biologically active native conformation of a protein is one of the fundamental challenges of structural biology. This problem remains yet unsolved mainly due to three factors: the partial knowledge of the effective free energy function that governs the folding process, the enormous size of the conformational space of a protein and, finally, the relatively small differences of energy between conformations, in particular, between the native one and the ones that make up the unfolded state.

Herein, we recall the importance of taking into account, in a detailed manner, the many interactions involved in the protein–folding problem (such as steric volume exclusion, Ramachandran forces, hydrogen bonds, weakly polar interactions, coulombic energy or hydrophobic attraction) and we propose a strategy to effectively construct a free energy function that, including the effects of the solvent, could be numerically tractable. We also describe the situation in which the native conformation has different covalent constraints than the unfolded state (such as disulfide linkages), and then the exact native structure can not be reached using the original free energy function. Finally, we discuss about the limits and the lacks from which suffer the simple models that we, physicist, love so much.

†) Correspondence to P. Echenique: `pnique@unizar.es`

PROLOGUE

(SELECTED MEMORIES OF J. L. ALONSO)

I became aware of Alberto Galindo in April 1965, when Professor Ortiz Fornaguera, who made a report on a scholarship project of mine for the *Fundación Juan March*, told me about him. The general ideas of my project fit in a good proportion, as Ortiz said in his report, to Bohm's remodeling of quantum mechanics.

At that time, I thought that there was nothing more beautiful than quantum mechanics and relativity, and I knew by heart Terradas and Ortiz's book on the subject [1]. Now, I think that it was, above all, my ignorance which made me prefer what I knew to some extent. But, after all, a first love is the most passionate one.

The thing is that, when Ortiz, for whom I felt a deep reverence, advised me to approach Alberto as the person who could best put my ideas in order, I thought it would be great if Alberto Galindo would agree to be my Ph.D. thesis director.

In October that same year, a couple of days after the Pilar fiestas, I was in Alberto's room in Zaragoza University, with my project about hidden variables under my arm. I was surprised when he immediately agreed to be my thesis director. Later, I learned that he was dead short of assistants, as in the *Junta de Energía Nuclear*, in Madrid, a Ph.D. course had been organized which would be attended by the best students in Madrid, Zaragoza, Valencia and Barcelona. That course remains unsurpassed, despite uncountable attempts at reproduction, and was crucial for the later development of Elementary Particle Physics in Spain. So, to my ego's abasement, it was not my project that affirmed Alberto's acceptance to be my thesis director. But even if Alberto thought that an assistant fallen from the sky was a present from the Virgin of the Pilar, it was not a bad start.

As usual, along his magisterial scientific career, Alberto worried at that time about subjects apparently as far apart from each other as the uniqueness of the position operator for relativistic systems [2] and the coupling of internal and space-time symmetries [3], both with large conceptual and mathematical content. This helped in assuring that he was not displeased that I devoted part of my time to pondering on the philosophical foundations of quantum mechanics, so closely related to the subject of hidden variables. I hope that my use of the term *philosophical* is understood in this context and does not rise any untimely debate.

According to my notes, my incursion into the subject ended when I told Alberto that pursuing such a goal meant to consider time as an operator. Too strong an assumption both for him and for myself.

In this short recall of my links with Alberto in the last near 40 years, I must leave out many things, such as the unforgettable time we spent in Orsay. I shall skip too the crucial role he played in the creation of the *Grupo Interuniversitario de Física Teórica (GIFT)*. I will only call the attention on his efforts to keep High Energy Physics at Zaragoza at a high level, first convincing Ángel Morales and Rafael

Núñez Lagos to exile in Zaragoza, where, at that time, Francisco Ynduráin still was a Professor, and later, along the years, favouring the collaboration of members of his department in Madrid with members of ours in Zaragoza. I presume that Guillermo García will speak at length about the Madrid-Zaragoza connection.

In any relatively important matter in which our department has been involved, Alberto has involved himself. He did it in 1997, when he took sides, together with some of us, dissuading the Government of Aragón from carrying on the project of Accelerator Assisted Fission. The taking of sides charged momentarily its toll, but Alberto is not the type man to support a project which he deems badly conceived, even if he is bound by a deep affection to its promoters.

More recently, in 2001, members of our department and members of the departments of Biochemistry and Molecular and Cellular Biology, Applied Mathematics and Condensed Matter of Zaragoza University started the Institute of Biocomputation and the Physics of Complex Systems (BIFI). From the first moment, not only did several members of the Theoretical Physics Department of Madrid Complutense University got involved, but Alberto himself supported us most decisively in the institutions in Aragón region requiring his advice. At the time of writing, BIFI has members in the Complutense, Carlos III, Extremadura, Elche and Granada universities and in the Institute of Science of Materials of Madrid (CSIC).

For some of us, broadening our scientific interest to biophysics is turning out to be a challenging experience. Our contribution to this book is in the realm of the protein-folding problem.

A protein is a biopolymer composed of monomer amino acidic *building blocks* (See Figure 1). These biological macromolecules fold at an amazingly fast rate, at least initially, after being synthesized in the ribosome, and eventually reach a well defined three-dimensional structure, known as *native structure*. It is in this particular conformation that each protein performs its specific biological task. By unfolding and refolding the ribonuclease protein *in vitro*, Anfinsen [4] showed that all the information needed to reach the final native structure is encoded in the sequence of amino acids, i.e., there is no cellular machinery needed to fold proteins (which is true, at least, for most of them). What we now call the *protein-folding problem* consists on the prediction of the native structure of proteins starting from the knowledge of the amino acid sequence and from the physical laws governing the interactions between atoms in the frame of statistical mechanics.

> I always feel like running away when any one begins to talk about proteids in my presence. In my youth I had a desire to attack these dragons, but now I am afraid of them. They are unresolved problems of chemistry; and let me add, they are likely to remain such for generations to come. Yet every one who knows anything about chemistry and physiology, knows that these proteids must be understood, before we can hope to have a clear conception of the chemical processes of the human body [5].
>
> – Professor Ramsen, Ira (1904) –

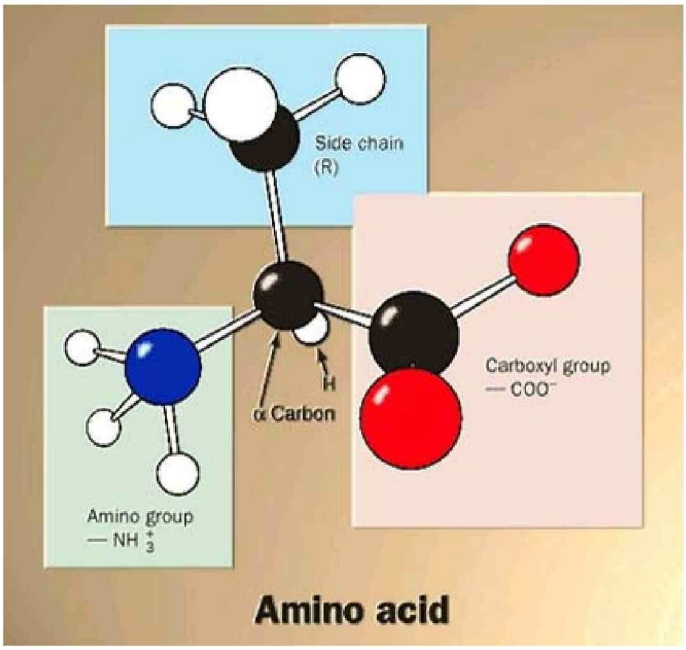

FIGURE 1. General schematic representation of an amino acid in its zwitterionic form, the sidechain part of the molecule is what distinguishes one amino acid from another. There are 20 naturally occurring amino acids with sidechains of differing complexity. The one depicted is one of the simplest chiral amino acids: the alanine.

At the time when Professor Ramsen threw down the gauntlet and challenged chemists to solving the proteinfolding problem, he did not realize that the solution was to perhaps come from a myriad of disciplines, not excluding those of theoretical and computational physics.

Our paper is organized as follows. In the next section we discuss mesoscopic organization in biological and non-biological matter. Afterwards, we describe the complexity of the proteins in terms of their energy landscape and we talk about the Flory isolated pair hypothesis in the context of these particular macromolecules. In other two sections we recall what one can learn from simple models and the necessity to go to detailed models if one aims at true predictive power. The next section is devoted to the water inclusion, and in the penultimate one we describe the two complementary approaches to the protein-folding problem, the *holistic* and the *reductionistic* one. Finally, in the last section, we discuss the confluence of these two views in our goal of a step-by-step construction of a detailed effective model with the final objective of capturing the essential characteristics that make the protein-folding process such an efficient and cooperative phenomenon.

INTRODUCTION

ARE WE TALKING PHYSICS?

In the year 2000, six *grand challenges* were identified and listed in the final report of the *Physics in a New Era* series by a high-level panel of the U.S. National Research Council (NRC), chaired by Thomas Applequist of Yale University.

They are as follows:

1. Developing quantum technologies
2. Understanding complex systems
3. Applying physics to biology
4. Creating new materials
5. Exploring the Universe
6. Unifying the forces of Nature

Regarding the challenge entitled "Applying physics to biology", the mathematician Stanislaw Ulam once redefined the problem by telling a biological physicist that he had finally understood the challenge: "Ask not what physics can do for biology; ask what biology can do for physics" [6].

Possibly, Stanislaw Ulam meant that biology provides excellent proving grounds for physics-based ideas about complex systems. But then one would be addressing the second challenge, "Understanding complex systems", rather than the third. One may then become suspicious that the interest of physicists, at least of theoretical physicists, lies not in biology itself but rather in "as-yet-undiscovered organizing principles that might be at work at the mesoscopic scales, intermediate between atomic and macroscopic dimensions, and the implications of their discovery for biology and physical science" [7]. In the case that the interest lies on mesoscopic organization, non-biological grounds like spin glasses or strongly correlated electron systems are possibly easier to code and more likely to allow new laws responsible for mesoscopic organization to be extracted, if they exist at all.

In this work, we shall approach a huge riddle, the high challenge for science that is one of the most central phenomena of biological processes: the protein-folding problem. We shall see that the uncertainties involved in it are so overwhelming that, if one's purpose is to find *new undiscovered organizing principles*, this problem does not seem to be the best place to look, at least at the present time. The unraveling of the mysteries of protein folding is, in fact, a clear example of a problem that fits into the third grand challenge of the NRC, "Applying physics to biology".

Then, if this third challenge does not interest theoretical physicists, who is it addressed to? Perhaps to computational quantum chemists. Perhaps it is such an interdisciplinary field that it places all disciplines *at the brim of a nervous breakdown*.

Yet, we physicists are expert at modeling nature, and applying physics to biology may mean a challenge to our modeling skills as impressive as the discovery of new organizing principles. The most important ingredients of a good model are its predictive power and its supplying us with a perception of the reasons for that power, i.e., with a hint of *what is responsible for what*. One might also add simplicity as an ingredient to good models. It has certainly been, up to now, one of the most attractive features of successful physical models. Now we know that the simplicity of the Hamiltonians used in physics is due to the proximity to fixed points in the parameter space of possible Hamiltonians. Behind that feature lies the fact that the interesting behaviour of traditional physical systems does not depend crucially on the details of the system. On the contrary, in complex systems, such as proteins, the interesting behaviour does depend crucially on the details.

If we perturb a protein *a little bit* (e.g. slightly altering the pH or substituting just one selected amino acid in the chain), the folding process may change dramatically and the biological activity of the protein may cease altogether. The existence of allosteric proteins (which drastically alter their shape and properties when they link a small regulating molecule like a vitamin) is a good example of this fact. We are mindful of biological functions associated with tiny structural details! Our very lives are at stake!

Even Professor Erwin Schrödinger himself would have agreed on this point. In a set of lectures given in Dublin, in 1943 [8], he stated the following:

> Every particular physiological process that we observe, either within the cell or in its interaction with the cell environment, appears —or appeared thirty years ago —to involve such enormous numbers of single atoms and single atomic processes that all the relevant laws of physics and physical chemistry would be safeguarded even under the very exacting demands of statistical physics in respect of large numbers [...] Today, we know that this opinion would have been a mistake. As we shall presently see, incredibly small groups of atoms, much too small to display exact statistical laws, do play a dominating role in the very orderly and lawful events within a living organism. They have control of the observable large-scale features which the organism acquires in the course of its development, they determine important characteristics of its functioning; and in all this very sharp and very strict biological laws are displayed.

To sum it up, if we intend to model a protein's behaviour, simplicity is best left aside and focus must be made on predictive power and the capability to enhance the perception of the reasons for that power. In contrast, if the desire is to identify yet undiscovered organizing principles that might be at work at mesoscopic scales, possibly there are other places to look, perhaps in systems that do not suffer from so much uncertainty.

PRINCIPLE OF MINIMUM FRUSTRATION

THE CORRECT FREE ENERGY FUNCTION SHOULD HELP IN SOLVING THE HUGE NUMERICAL PROBLEMS

Today, we know that the two problematic paradoxes of protein science are neither problematic nor paradoxical [9]

The first one of them is known as the *blind-watchmaker paradox* [9,10] and it equates the vastness of the sequence space of polypeptides with the impossibility of ever finding a protein-like sequence.

For a chain of, say, 100 natural amino acids, there are $20^{100} \simeq 10^{130}$ possible sequences. Therefore, the probability of observing the sequence of a particular protein is negligible. This problem has been regarded as impossible to solve by creationists, who appeal to divine intervention and has been circumvented by evolutionists through the mechanisms of natural selection.

The fact is that there is not really a problem with the numbers. It has been shown by statistical modeling [11] that the probability of pulling out from a *soup* of random amino-acid sequences a particular one that folds to the same structure and performs the same biological function as protein A is much more than 10^{-130} (which would be the probability of extracting *exactly* the sequence of A). It turns out to be more of the order of $10^{-20} - 10^{-10}$ [9], which is a probability that, in spite of being still small, can be easily overcome by natural selection. Besides, the probability of pulling out a sequence that folds to any well defined native structure, not just that of A, is even larger.

So it seems that there is an enormous degeneracy in sequence space, i.e., differences in sequence do not necessarily imply differences in biological function or in fold. But, how can one easily explain this degeneracy? It is known that the particular three-dimensional structure of a protein is determined to a great extent by sidechain forces. In the folding process, these sidechain forces cooperate with backbone forces to reach the stable native conformation in $10^{-5} - 10$ seconds after the sequence is synthesized at the ribosome. Although the true balance between sidechain and backbone forces is not yet known, it is impossible to fold a protein against the sidechain forces. The backbone forces are essentially present with the same magnitude irrespectively of the particular sequence and it is the more sequence-specific sidechain forces that help direct the fold. However, simple exact models [12] show that the precise information of the sequence is, most of the times, redundant; it has been found that the fold is primarily determined by the sequence written in a two-letter alphabet rather than in the natural twenty-letter alphabet. One can classify the amino acids into two categories regarding their affinity for water: hydrophobic (H) and polar (P). Using this code, it is found that, if a certain sequence does fold, the sequence obtained by interchanging one hydrophobic amino acid for another hydrophobic amino acid (analogously for polar ones) will fold with a very high probability to a very similar structure. Thus, the essential features of the full $20^{100} \simeq 10^{130}$ sequences space remain in the smaller space of the sequences

written in the HP alphabet, which contains *only* $2^{100} \simeq 10^{30}$ elements. Moreover, experiments [13] show that only about $^1/_3$ of the residues are crucial for folding, specifically those that define the hydrophobic core of the protein. Adding this fact to the preceding one, we see that the real search for protein structure takes place in a space whose size is closer to $2^{100/3} \simeq 10^{10}$ than to the overwhelming 10^{130} first proposed, the remaining 120 orders of magnitude are highly degenerated.

As it was stated, 10^{10} is still a huge number, but it is affordable for natural selection to search in such a space. Natural selection is a *blind watchmaker* [10] that selects, among these 10^{10} potential protein sequences, the true ones through a partially directed process that, concurrently, involves considerable random choice among alternatives.

Once Nature has selected a protein for its capability to fold and to perform a certain biological function, one may wonder how this molecule can find the native structure in such a short time, considering the many degrees of freedom it has and the frustration among the different interactions. This question was, for years, regarded as a paradox: the *Levinthal paradox*.

It was first stated in a talk entitled "How to fold graciously" given by Cyrus Levinthal in 1969 and noted down by A. Rawitch [14]. The paradox states that, if in the course of folding a protein is required to sample all possible conformations (a hypothesis that ignores completely thermodynamics and statistical mechanics) and the conformation of a given residue is independent of the conformations of the rest (which is also false), then the protein will never fold to its native structure.

For example, let us asume that each residue of a chain of 100 amino acids can take up to 10 different conformations on average (typically, there are 9 relevant backbone conformations and a variable number of sidechain ones, but let us use 10 for the sake of simplicity). This makes a total of 10^{100} different conformations for the chain. If the conformations were sampled in the shortest possible time ($\sim 10^{-13}$ s, i.e., the time required for a single molecular vibration), one would need about 10^{77} years to sample all the conformational space. This result implies that protein folding cannot be a completely random trial-and-error process, as one already could have imagined by taking into account the laws of thermodynamics and statistical mechanics. Maybe, calling this problem a *paradox* is too much (in fact, Levinthal did not use this word and, just after stating the problem, he explains a possible solution to it). It is clear that the size of the conformational space of a polypeptidic chain is astronomically big (even if one takes into account the fact that the conformations of different residues are not independent) and we must explain how the system can navigate through it from the unfolded state to the native conformation in such short time.

Folding is not a general property of heteropolymers. Heteropolymers, due to their many degrees of freedom and the many geometric constraints among them, are said to present a great *frustration*, that is, there is not a single conformation of the chain which optimizes all the interactions at the same time. In any conformation, the different interactions are conflicting, i.e., *frustrated*. In polypeptides, as in many heteropolymers, frustration is mainly due to chain connectivity between monomers

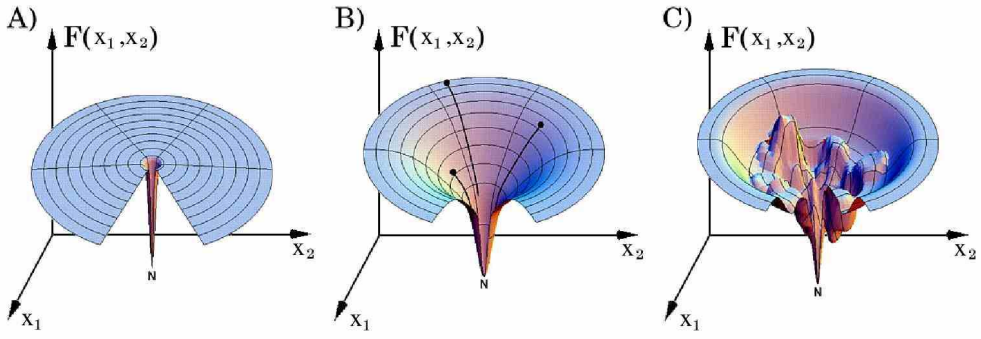

FIGURE 2. Possible energy landscapes of a protein. The conformational space is asumed to be two-dimensional, the degrees of freedom being x_1 and x_2. The free energy $F(x_1, x_2)$ is a function of these two variables, which are internal degrees of freedom of the molecule. The degrees of freedom of the solvent have been *integrated out*. **A)** The energy landscape as it would be if the Levinthal Paradox was a real problem. **B)** A smooth funneled energy landscape. **C)** A more realistic partially rugged funnel. N stands for native state and it is assumed here to be the global minimum. (Figure taken from Dill K. A., Chan H. S., *Nat. Struct. Biol.* **4**, 1 (1997) and somewhat modified)

with opposite affinities to neighbours and/or environment. This leads to a rugged energy landscape with many low-energy states, high barriers, strong traps, etc.; up to a certain degree, a landscape similar to that of spin glasses. Hence, to reach a stable well-defined three-dimensional native structure, a protein molecule cannot have a totally rough landscape. On average, its native conformation must be more stabilizing than would be expected for a polypeptide of random sequence [15].

Bryngelson and Wolynes [16,17] have termed this fewer conflicting interactions, than typically expected, as the *principle of minimal frustration*. This takes us to a natural definition of a *protein* (opposed to a general *polypeptide*): a *protein* is a polypeptidic chain whose sequence has been naturally selected to satisfy the principle of minimal frustration. Such a molecule is allowed to rapidly fold by trading conformational entropy for internal energy as it moves down a landscape that is funneled (by virtue of the principle of minimal frustration) towards the native structure (see Figure 2).

Recent theoretical and experimental evidence [18] fully support this folding scenario, i.e., biological proteins are only minimally frustrated and the folded conformation can be reached by one of a large number of paths [19].

In the context of protein folding, if one puts aside the explicit solvent and agrees to treat it implicitly (which, given the present power of computers and understanding of solvation at the molecular level, is a must), the central physical object is the internal free energy $F(\vec{x})$ of the molecule (see Figure 2). This energy function depends only on $\vec{x}$, the degrees of freedom of the polypeptidic chain. The degrees of freedom of the solvent molecules have been *integrated out* and that is the reason

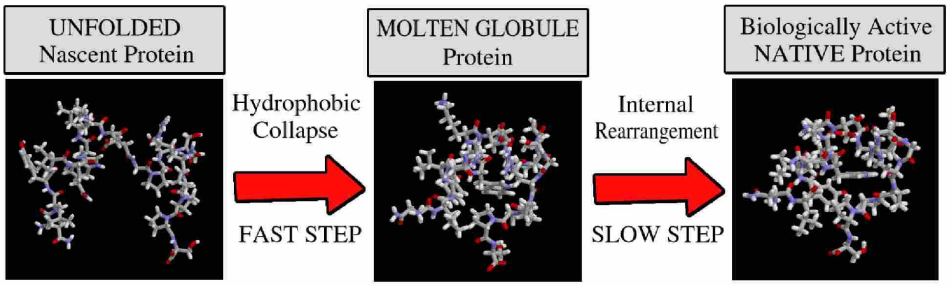

FIGURE 3. A schematic illustration of the multistep process of protein folding as compressed into two major steps. Since the first step is *fast* (a few miliseconds) it may correspond to a rolling down process from the *highland area* of the potential free energy hypersurface to the *lowland area*. Finding the final native conformation on the *lowland area* is a considerably slower process.

of calling this energy *free*, meaning that it depends parametrically on the temperature. It contains the entropy of water, but not the conformational entropy of the polypeptide (that is the reason of noting it as F and not as G, we reserve the letter G for the total thermodynamic free energy of the system, which only depends on the temperature and the pressure and not on the internal degrees of freedom). Of course, if one works at constant temperature, $F(\vec{x})$ can be regarded as a normal potential energy and a single conformation of the chain can be considered as a unique microstate of the system in the mean field of the solvent rather than a truly dynamical object. Finally, if one refers specifically to the *graphical* features of $F(\vec{x})$, it is usual to use the term *energy landscape*.

We take into account that it is an experimental fact that the native state consists of practically only one conformation or, more precisely, a resonance of energetically-closely-spaced minima. During the last quarter of the 20th century a hypothesis emerged, which eventually became a dogma, namely, that the native conformation must be the global minimum of $F(\vec{x})$. Some theoreticians agree about this point [4,20], which in the context of protein folding is referred to as the *thermodynamic hypothesis*.

Currently, science has a better understanding of the mechanism of protein folding: it consists of two complex steps interconnecting three phases [21–26], as shown schematically in Figure 3.

In the first step, the secondary-structural elements (e.g. helices, β-pleated sheets, etc.) are formed and the hydrophobic sidechains come into each other's vicinity, forming a hydrophobic core. We will refer to this state, which has the majority of the secondary-structural elements developed but whose tertiary structure is not completely formed, as the *molten globule* state [27,28].

In the second step, the hydrophobic core is reorganized and a slow search for the native structure begins. For some proteins, also high-energy bonds are formed such as disulfide linkages (their formation is $\sim 8\ kcal \cdot mol^{-1}$ unfavourable) and

trans $\leftrightarrow$ *cis* isomerization of proline peptide bonds occurs. Formation of the disulfide linkages is aided *in vivo* by *protein disulfide isomerases* and the cis proline peptide bond isomerization is catalysed by *peptidyl-proly-cis/trans isomerases* [29,30]. When this happens, the high-energy bond formations establish new constraints in the system, favouring to populate a conformation that, in the former *unconstrained* protein, would have been thermodynamically less stable. This free-energy minimum of the new system (the protein with redefined constraints) is the native state and it is very proximate in the phase space to the conformations that made up the molten-globule state. Of course, when no constraints redefinition occurs, the stability of the native state is enough for the system to populate it preferentially.

In the view of the aforementioned, it would seem reasonable to break up protein-folding studies into *two steps*, along the lines suggested by Figure 3 [27,28]. What is typically being pursued with existing force fields, irrespectively whether by Molecular Dynamics or by Monte Carlo search, would fit in the *first step*, that is, in finding the conformations that make up the molten globule state of the protein. The aim of *walking the last part of the road* to the native state is still numerically unfeasible. Methods must be designed in order to simulate this slow process or a new generation of faster computers must be waited for. In the cases in which the internal free-energy function F does not include the constraints present in the native state, i.e., the disulfide linkages and the new value of the ω angles of some prolines, whose description may be thought as a redefinition of the geometrical constraints of the system, we need also to build new software applications. *In vivo*, both the formation of the disulfide linkage as well as the *trans* $\leftrightarrow$ *cis* isomerization of the peptide bond, associated with the proline nitrogen, are enzyme-catalysed. However, the process could be mimicked without the involvement of the enzymes, as the thermodynamics of the process is the same for both uncatalysed and catalysed reactions. To the best of our knowledge, no systematic work has been carried out to accomplish the construction of a force field aimed at simulating these processes, but some preliminary research has been done in the field of disulfide linkages formation [32] as well as *trans* $\leftrightarrow$ *cis* isomerization of peptide bonds including both glycine and proline [33,34].

Under these conditions, a reasonable strategy to find the molten globule, would be first to try to write a free energy potential $F(\vec{x})$ that incorporates, in the simplest possible way, all the interactions that play an important part in the folding process. Then, the problem is to locate the *lowland area* of $F(\vec{x})$, which depends on a large number of variables. Finding minima is a classical problem of numerical analysis and a number of strategies are possible (even *genetic* algorithms!). We will consider algorithms closely related to statistical mechanics and based on Monte Carlo simulations, which are efficient when a large number of degrees of freedom are involved and present an important advantage: it is also possible to simulate the behaviour of the system as a function of temperature. To find minima using Monte Carlo methods, one must replace the traditional approach (whose aim is the generation of a Boltzmann ensemble of conformations) by an algorithm designed to rapidly identify the lowland area of $F(\vec{x})$ [35]. If this internal free energy function

properly describes the underlying physics, the giant size of the conformational space is not a hindrance, as $F(\vec{x})$ would be funneled towards the molten-globule state and the efficiency of the algorithm would be greatly increased.

STERIC FORCES

THE IMPORTANCE OF PAYING THE FAIR PRICE (ENTROPIC)

To reach the more accurate internal free energy function mentioned in the preceding section, it is desirable to find a set of interactions such that, for the particular protein system and at normal biological conditions (temperature, pH, salt concentration, etc.), one can asume that factorization is possible, i.e., that the total energy can be written as a sum of different components, one for each relevant interaction. This possibility is dually convenient: on the one hand, data to write each of the energy terms can be extracted from experimental or *ab initio* results carried out on small simple systems in which only a subset of the complete set of interactions is present and, on the other hand, when testing the internal free energy function, one can *switch on* or *off* any of the interactions in order to gain insight about *what is responsible for what*. However, factorization is not always possible, because the different *interactions* in which one tries to divide the problem are actually part of the same fundamental forces and, in this way, a residual coupling may remain. Moreover, as we have already mentioned, many microscopic degrees of freedom must be *integrated out* if one wants to ever solve the problem. This elimination process causes the remaining degrees of freedom to couple. The example of water is illuminating: before the integration of the solvent degrees of freedom, the electrostatic contributions as a function of atomic positions can be considered practically pairwise; after the integration, however, the electrostatic interaction energy of a pair of atoms depends on the positions of the rest.

In this section we shall talk first about one of the important interactions that play a role in the folding of proteins, the steric forces beyond nearest neighbours. We will dig into an example of how, if one does not account in detail for these forces, the entropic price that must be payed as the protein rolls down the real funneled landscape becomes so distorted that there is no hope of ever getting to a useful $F(\vec{x})$. Obviously, the entropy of the unfolded state is a determinant quantity in this discussion, because, the entropy of the molten-globule state being very small, it is just the entropy of the unfolded state that the favourable interactions must overcome in order to fold the chain.

Along many years, the original Zimm-Bragg [37] and Lifson-Roig [38] helix-coil theories have been greatly extended [39] with the aim of codifying experimental data of peptide helices in solution [40,41]. These theories have been markedly influenced by simplifying *Flory's isolated-pair hypothesis* [42], which states, in the context of proteins, that each Ramachandran pair (Φ, Ψ) [43] in the peptide backbone (See Figure 4) is independent. It is rather obvious that this simplification suffer greatly from over- or underestimations in order to guarantee predictive power to the models

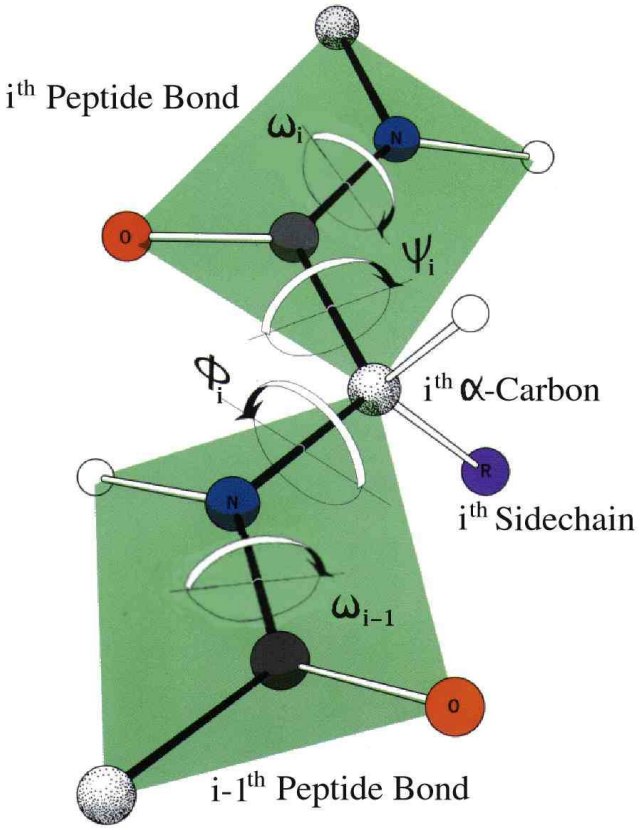

FIGURE 4. Angular degrees of freedom of the polypeptidic chain. Ramachandran angles Φ_i and Ψ_i are the most flexible degrees of freedom, the peptide bond angle ω_i is more rigid because of the structure of the π resonant bond. (Figure taken from Voet D. and Voet J. G., *Biochemistry*, 2nd Ed., John Wiley & Sons, 1995, and somewhat modified)

that incorporate it, however its utility in the codification of peptide helices data has been substantial [39–41]. In fact, it follows from the hypothesis that *local* structure transitions are ruled out as a possible source of cooperativity in protein folding [44]; the entropic price is simply too high for short polypeptide backbones to preferentially populate a small set of highly similar conformations.

In a recent paper, Pappu, Srinivasan and Rose [45] have proved that, in the aforementioned helix-coil theories, the volume of the coil region is much smaller than predicted by the isolated pair hypothesis, thus lowering the entropic barrier for helix formation. In a real protein, most conceivable states allowed by Flory's hypothesis are not really accessible and steric effects beyond nearest neighbours bias the unfolded molecule towards organized structure [46]. The *hard-sphere* model may be an effective treatment of the most relevant steric repulsive interactions beyond nearest neighbours [45].

Another important component of the free energy function $F(\vec{x})$ is the one that accounts for the backbone interactions between neighbouring peptidic planes (See Figure 4). The rotation around the Ramachandran angles (Φ, Ψ) is hindered by clashes between atoms around the α-carbon and by strains in the electronic clouds involved in the bonds. However, not all the bonds in the backbone can rotate; the π orbitals that partially contribute to the peptidic bond severely limit the rotation around it and that is the reason for the peptidic bond being nearly planar. This fortunately allows us to decouple the potential energy surface (PES) that describes the rotation around the Ramachandran bonds of one particular residue from the PES of its neighbours, which has permitted to render useful the *ab initio* quantum mechanical calculations on dipeptides aimed to extract the one-residue PESs *in vacuo* [47–49]. This established works show that the energy differences in the (Φ, Ψ) space can be very large: of the order of $\sim 10\ kcal \cdot mol^{-1}$. As this number is of the order of the total folding energy, it is therefore clear that there are zones of the Ramachandran space which are almost energetically prohibited (these facts do not seem to change very much with the inclusion of implicit solvent [50,51]). This component of the free energy is crucial to be accurately implemented, as any error in the relative energies between different values of (Φ, Ψ) will be repeated as many times as the number of residues of the chain.

Two other relevant *in vacuo* interactions that remain to be discussed are the coulombic forces (the portion of them that has not been taken into account in the two preceding components) and the hydrogen bonds. In fact, it is not clear whether hydrogen bonds must be considered as being something essentially different to a dipole-dipole interaction, however, it has been demonstrated that the specific characteristics of these interactions, i.e. short range, directionality and molecular orbital symmetry, are crucial to stabilise important structures of the proteins, such as the α-helices [52]. Of course, when the effect of water is added to this energy (a topic discussed further on), some details of these interactions are modified. However, the directionality and energetic influence of the hydrogen bonds between peptidic planes remains practically intact, as their quantum-mechanical *ab initio* characterization may be decoupled from solvent effects [53].

SIMPLE MODELS

WHAT CAN WE LEARN FROM THEM?

As aforementioned, in molecular biophysics in general and in the protein-folding problem in particular, it is not possible to design accurate and precise simple models, that is, models that can quantitatively predict at least some aspects of the behaviour of real systems but, at the same time, contain a reasonably small number of parameters.

Yet, the contributions of simple models have been qualitatively relevant in the few last decades to shed light on the general principles and mechanisms, as well as to seed new ideas concerning the type of physics involved. These contributions

have allowed the ruling out of some views that had been uncritically adopted from the domain of small-molecule chemistry. The ideas of a folding funnel and of parallel pathways in protein-folding dynamics stem from the results of simple physical models, which also contribute in proposing possible scenarios for molecular evolution, in describing the denaturation of DNA, in predicting the secondary structure of RNA and in a number of other domains of molecular biophysics [12,54–58].

Among the studies aimed at increasing understanding of the *common characteristics* of globular proteins, those that deserve special attention are the ones that show these characteristics arising from considerations of symmetry and geometry associated with the polypeptide chain and that only a limited number of folds arise from geometrical constraints imposed by sterics and hydrogen bonds [59]. Any more sophisticated model that wants to take a step further in predicting the details of the folding of a particular protein will not be able to contradict this type of analysis. It must go further by incorporating details of the steric forces, the hydrogen bonds and at the same time of long-range coulombic interactions and others of entropic origin that, cooperating with the former ones, stabilize one appropriate fold among those that symmetry and geometry have previously selected.

DETAILED MODELS

SOME OF THEIR INGREDIENTS AND DIFFICULTIES

A detailed model does not necessarily have to include all the atoms of the molecule, nor does any one atom require treatment at the same level of coarseness in all the interactions in which it plays a role. For example, as mentioned, a hard-sphere description of an atom may be appropriate to account for its repulsive steric interactions with other atoms in the amino-acid chain, on the contrary, it is completely inadequate to reproduce the PES associated with the Ramachandran degrees of freedom. With such a description, all allowed (Φ, Ψ) pairs would have the same energy, a result that *ab initio* studies [47–49] prove absolutely false.

Clearly a detailed model of any physical system need not to include all the degrees of freedom, although it does have to include those that are considered the largest contributors to the partition function of the system or, alternatively, to an equivalent effective system, if a part of the system (we are thinking about, for example, the part that corresponds to the water molecules in a protein system) has not explicitly been taken into account. Generally speaking, the contribution to the partition function of the vibrational degrees of freedom is negligible compared to that of the rotational ones at 300 K (See table 11.2 in ref. [60]), however, not all the rotational degrees of freedom themselves play an equivalent role. In the case of proteins, the dihedral angles around the $N - C_\alpha$ bond (Φ) and around the $C_\alpha - C'$ bond (Ψ) determine the *main-chain* or *backbone* geometry, from which stem the different sidechains and must be undoubtly incorporated to any model, as they define the secondary and tertiary structure. On the contrary, the rotational degrees of freedom of the sidechains, which are responsible for the different conformations

of these groups, known as *rotamers*, are certainly of secondary importance when compared to the Ramachandran angles (Φ, Ψ) and may be implemented in a more simplified way.

Once the relevant degrees of freedom of the system are known, in order to have any opportunity to find the molten globule state, we must write a free energy function detailed enough. As it has been already explained, the relevant interactions that must be included are: steric constraints [42,43,61,62], hydrogen bonds [52,63], Ramachandran PESs [47–49], coulombic interactions and hydrophobicity [64].

On the topic of the different approaches to the protein-folding problem, R. A. Abagyan [65] has classified them in a way that is much liked by the authors of this article. According to Abagyan, we can divide the scientists that try to predict the native structure of the proteins into three categories: *dynamicists* [66–68], *minimalists* [69–74] and *synthesists* [20,75–77]. In Abagyan's own words:

> Dynamicists believe that sufficiently long simulations of a quasi-continuous trajectory of molecular dynamics of atomic models *in vacuo* or in water will solve the problem using new generations of computers, code parallelization, and optimized simulation techniques. Minimalists, unwilling to play power games and too impatient to wait until new generations of processors cover the next mile of a hundred mile road, simplify the system by using a reduced atomic representation or a lattice, inventing a potential and then enjoying the luxury of always finding the global minimum of their energies as well as most of the other possible states for a chain of up to a hundred simplified residues [73,74]. The third school shares the impatience of minimalists, yet resists the temptation to use simple models since it appears that accuracy is a pivotal issue. Synthesists focus on the development of algorithms to replace molecular dynamics as a generator of conformational changes and the design of methods of energy calculation which combine accuracy and speed.

Abagyan includes himself in the group of the synthesists and gives some more detailed characteristics of this approach. Since the synthesists are, of the three categories, the scientists which we also find ourselves in better agreement with, we quote again the words of Abagyan to describe some of our shared ideas. Although points 1 and 3 require further qualifying.

> Let us list some of the ideas and assumptions of the synthesists, including the author, which the following review is based upon:
>
> 1. Oscillations of bonds lengths and bond angles are not essential for protein structure prediction and some of these degrees of freedom are not even excited at room temperature. Therefore, using torsion angle space instead of Cartesian coordinate space is highly preferable since it reduces the dimensionality of the problem by a factor of 7, eliminates fast oscillations and smoothens the energy landscape.

2. A continous molecular dynamics trajectory is not really necessary for structure prediction. The optimal structure can be found by a global optimization algorithm making much larger steps.
3. Explicit consideration of water molecules can also be sacrificed in simulations of folding for the following reasons: (i) too many additional degrees of freedom; (ii) a really large box is necessary because of the long-range nature of the electrostatic interactions; and (iii) the relaxation times of water molecules after a large conformational change is prohibitively long. Concurrently, the solvation effect can be evaluated by continuous approximations more efficiently and, potentially, more accurately.
4. The correct conformation and an enormous number of altenative conformations of a polypeptide chain may have very close energies. A high accuracy of energy calculations is absolutely essential to recognize the correct answer.

To conclude this section, an implicit *bona fide* assumption has been held, namely, that one will be able to write a physical $F(\vec{x})$ that will produce a funneled landscape and that numerical algorithms to find minima will smoothly roll down the funnel in an acceptably short time. A more realistic scenario forces one to realize that a free-energy function that is slightly different from the real one combined with the high dimensionality of the phase space could cause significant numerical problems. Even if there is no strong frustration phenomena, to travel around (almost) all phase space is very time consuming (ergodicity). If we do not study large regions, we will not be able to guarantee that any lowland area found is the lowest one. When a large number of degrees of freedom are considered and one wants to study a true protein-folding problem, Monte Carlo simulations are the most efficient algorithms to obtain minima. However, continued improvement of these algorithms is needed in order to speed up simulations and to facilitate more efficient parallelization.

THE WATER

WHERE THE WHOLE STORY TAKES PLACE

The actual shape of the free-energy fuction of the protein $F(\vec{x})$ is determined to a large extent by the fact that the folding takes place in a particular environment: an aqueous one. Let us examine the basis for this.

Although the Ramachandran PESs and the steric volume exclusion have their origins in strong *electronic* interactions, the weak intramolecular bonds in the protein have a crucial role in the preference of the differing conformations of the polypeptide chain. The energy of formation of the weak bonds lies in the interval of $1-5$ $kcal{\cdot}mol^{-1}$ (See the book by Dill and Bromberg [60] for an inspired discussion about the hierarchy of the energies of the different molecular interactions). At room temperature, the average kinetic energy of a water molecule is $RT{\simeq}0.6\ kcal\cdot mol^{-1}$,

providing the numerous water molecules with sufficient energy to break these weak bonds. Therefore, at physiological temperatures ($\sim 37^oC$), the weak bonds have, on the one hand, a transitory existence, although their joint action stabilize the relevant structures of the chain and, on the other hand, they are strongly dependent on the solvent. Moreover, the so called hydrophobic effect, which is an entropy-driven force, is genuinely characteristic of the solvent.

It is clear that, for a certain interaction to be relevant in the determination of the molten-globule state of a protein, it is a necessary condition that it changes the order of the different conformations on the energy axis determined by the rest of the interactions. The inclusion of water does change this order, even at such a fundamental level as the one-residue Ramachandran PESs [50,51] and hence why one is forced to take it into account. Of course, an interaction that does not change the energetic order of the conformations may still have an effect if one makes a molecular dynamics simulation, as the rates at which the transitions occur depend on the differences of energy and not only on their order. However, if one wants only to minimize the energy and to characterise the global minimum, although such an interaction may affect the convergence rate of the algorithms used, it will not change the fundamental features of the methods. The same criteria may be applied to the approximations used to compute the different interactions [78], i.e., one must check whether or not the particular approximation changes the energetic ordering of the conformations.

Regarding the question of how to implement the influence of water [79], it seems reasonable to assume that the effects due to the discreteness of the solvent are related only to the first layer of molecules around the solute. This first shell of molecules (the primary solvation layer) will be responsible for the entropic forces associated with a perturbation of the hydrogen-bond network and with the reduction of the phase space of water, which has traditionally been termed the *hydrophobic effect* [80–82]. To take into account these forces and also those of enthalpic origin, appearing when the primary solvation layer is confronted with a polar atom, it is common practice to write an energy proportional to the *Solvent Accessible Surface Area (SASA)* [83,84] of each atom. Beyond the first layer, water may be regarded as a linear isotropic and homogeneous dielectric medium. Neither in one zone nor in the other can one as yet afford the computational expense of treating water explicitly at the molecular level. The increase in numerical complexity would render the simulations near impossible at the present time, as we have already remarked when quoting R. A. Abagyan on the topic.

The rapid calculation of the energy of a conformation is a central issue in protein folding, and the treatment of water remains a problem since even the most approximated methods are relatively slow. To illustrate, we mention the molecular dynamics simulations conducted by V. Pande *et al.* [85]. These lengthy calculations have been performed with the Jorgensen's optimized parameters for liquid simulations (OPLS) force field [86] and, to account for the solvent, an approximated method called GB/SA [87], which not only avoids using explicit water but also approximates the Poisson equation for the dielectric region. The inclusion of explicit

water molecules would raise the simulation times by approximately three or four orders of magnitude.

Pande's research group has used distributed computing techniques and a *super-cluster* of thousands of computer processors around the world to address important questions regarding the β-hairpin folding (a fragment of a G protein) of only 16 residues. Although the results obtained from this massive simulation are very relevant, they disagree with a proposed mechanism [88,89] in which folding initiates at the turn of the β-hairpin and propagates downwards in a *zipper-like* way. Even if this disagreement is possibly due to the model used [88,89], we quote a revealing comment in the article by Pande: "The results of our study generally speak in favour of using the OPLS potential set for studying of protein fragments and small proteins. *However, the possibility that this set is not perfectly suitable for the direct folding of the β-hairpin remains.*"

This assertion shows that the design of a truly physical free-energy function is a more fundamental, and certainly prior, issue than the numerical problems associated with the simulations. This task is a challenge to our modeling capacity that could be to the physicists' liking.

THE TWO SIDES OF THE COIN

HOLISTIC AND REDUCTIONISTIC APPROACHES TO PROTEIN-FOLDING

There are two types of approaches in all scientific problem: *holistic* and *reductionistic*. Clearly, the field of protein folding is no exception! They consist, respectively, of the following:

1. Study the whole problem even if one must give up accuracy by using a non-rigorous method.

2. Study a small fraction of the whole problem, using a rigorous method. This is done in order to solve the whole problem more accurately.

What we have discussed so far is mostly the holistic approach, in which we take the whole problem and try to engineer a solution for it.

In contrast to that, there is the reductionistic approach, whereby one studies peptide folding, rather than protein folding, using first principle methods such as *ab initio* Hartree-Fock or Density Functional Theory with Gaussian split-valence basis sets of different sizes in order to solve the many electron Schrödinger equation. The thesis is that any understanding we might obtain from peptides will help us to understand protein folding. This has been discussed at some length recently [90].

Of course, the key question now is how long a polypeptide chain is called a *peptide* and how long it is called a *protein* (or a *polypeptide* if it does not fold to a well defined three-dimensional structure). Perhaps the border line between the two fields lies around 30 residues in the chain.

The 1990s have seen a lot of development on this [91,92]. Several single-amino-acid diamides have been studied computationally and, by the dawn of the 21st

century, all naturally ocurring amino acids (i.e. those with DNA codons) have been studied at least in an exploratory way [93]. The full conformational space of the dialanine dipeptide has been explored [94–96] and certain motifs of tri- and tetrapeptides have also been studied in a preliminary way. At that time, these achievements were possible by pushing the computational limits. Today, even larger tetrapeptides can be studied relatively easily [97].

Before the turn of the millenium, HF/3-21G calculations were carried out on a helix containing 12 alanine units. Only the 8 alanine unit has been studied further, at the DFT level, using a larger basis set and with solvation included [51]. We may hazard a guess that, within a couple of years (i.e. by the years 2006-2008), a helix containing 24 alanine residues will be computable at a similar (or even more accurate) level of theory. This particular length is very relevant, as it is the length of the section of the transmembrane proteins passing through the cellular wall [98]. Soon thereafter, a 30 residue polypeptide will be within the realm of the possibilities. Such a long polypeptide chain is no longer as flexible as the short peptides and it already posseses some of the intrinsic properties of proteins.

As a reminder, we should note that so far nobody explored the full conformational space of trialanine, which may have up to $9^3 = 729$ backbone conformers [99,100]. A tripeptide can offer an important structural feature, namely, that the central residue has a nearest neighbour at both sides. Similarly, tetrapeptides, such as tetraalanine, can offer a situation where a β-turn has nearest neighbours at both ends, however, they may exhibit up to $9^4 = 6561$ backbone conformations, which presently represent a computational hindrance.

It is hoped that, when these territories are covered, involving a number of different sequences of amino acid residues, not only alanine, a deeper insight will emerge and the computed data that support this insight will be useful for those dedicated to the holistic approach.

WHAT DOES THE FUTURE HOLD?

USING HIGH-LEVEL AB INTIO QUANTUM MOLECULAR COMPUTATIONS

In the design of simple models, one must recognise that the *future* always arrives sooner than one is prepared to give up the *present*. The *present* becomes the *past* and challenges that were put aside as being those of future work, soon require immediate attention.

The protein-folding problem posesses a number of these *future* challenges, including the treatment of explicit solvating particles (particularly the primary solvation layer), inclusion and evaluation of all thermodynamic parameters, analysis and understanding of the bases of molecular orbital symmetries and overlaps, in-depth systematic conformational analysis, as well as required compromises in the thoroughness of the theoretical treatment of the model peptide systems.

An ongoing reductionist (micro-molecular) *correction* to more holistic (macromolecular) models may be a promising route to more accurate development of

simple protein models [90].

One therefore requires an effective and analytical description of the whole problem, including therein a numeric definition of the relative spatial orientation of all constituent atomic nuclei [93]. Modular in design, of objective numerical efficiency (for ease of data set incorporation into novel Unix, Linux, Free-BSD and Perl- or Shell-script design and *vice-versa*), such a standardised nomenclature and analytic treatment may provide *smart data sets* to advance the solution. Far outperforming *smart programs* and *smart* computational architecture, refinement in the definition itself of the structural, conformational and energetic phase space provides an allowance (place holder) in an analytic solution, for any observable generated by the operators sampling the (wave)functions modeled.

The generation of accurate (wave)functions for the model becomes the focus of reductionist arguments, whereby the treatment of the electron density and correlation, thermodynamic and nuclear terms, approximation levels and the mathematical description employed to describe the physical space in relative proximity to the nuclei (the *basis set*), must all be of sufficient accuracy to account for all directing forces in folding.

The theoretical methods and level of theory employed in the modelling of polypeptides must therefore be calibrated to highly deconvoluted and highly resolved experimental data [101–105]. In-depth analysis of experimental neutron and x-ray diffraction (XRD), ultraviolet spectroscopy (UV), mass-spectrometry (MS), electron-spin resonance (ESR), circular dichroism (CD), infrared (IR), nuclear magnetic resonance (NMR) experimental results (among many others) must be incorporated in the derivation of an accurate and tractable energetic partition function.

As each residue in the polypeptide chain has its own unique structural and electronic contribution to the total evaluation, it may be prudent to provide, at least as a *present* consideration and theoretical exercise, the inclusion of subdivisions in the H (hydrophobic) and P (polar) residue distinction proposed before.

For example, the Phenylalanine (Phe) residue is commonly treated as being hydrophobic, yet it displays polar behaviour through interactions of its aromatic sidechain ring with polarised atoms. Aromatic-amide interactions are well established in the literature, among many other *weakly* polar intra-molecular ones [106–113].

A number of the stable Phe conformations, including the global minimum, show such *weak* interactions to be important in. The peptide bond also shows differences to its *rigid and planar* description, with deviations of $10 - 15^o$ being common, even for the global minumum [104,114,115]. The impact on the resonance structures of the peptide bond tautomer is not trivial and results in large differences in the polarisation of the consituent carbonyl ($C = O$) and amidic ($N - H$) elements. As these are the atoms responsible for the majority of intra-molecular interactions and hydrogen bonds, the peptide nitrogen's deviation from planarity could become quite important in the folding process. Proline demands even further scrutiny, as it shows a relatively large population of the cis-isomer in proteins [116–118].

The H (hydrophobic) and P (polar) categories could be subdivided into glycyl

TABLE 1. Subdivision of H and P categories for naturally occurring amino acids

H		P	
${}^{g}H$	Gly	${}^{Ar}P$	Tyr, Phe, Trp
${}^{s}H$	Ala, Val	${}^{o}P$	Ser, Thr
${}^{l}H$	Leu, Ile	${}^{s}P$	Cys, Met
${}^{Pro}H$	Pro	${}^{n}P$	Asn, Gln
		${}^{+}P$	His, Lys, Arg
		${}^{-}P$	Asp, Glu

(${}^{g}H$), small (${}^{s}H$), large (${}^{l}H$), prolyl (${}^{Pro}H$), aromatic (${}^{Ar}P$), oxygen- sulfur- and nitrogen-containing (${}^{o}P$, ${}^{s}P$ and ${}^{n}P$, respectively), cationic (${}^{+}P$) and anionic (${}^{-}P$) subcategories. The two last will be formed by protonation of His, Lys and Arg and by deprotonation of Asp and Glu, respectively (See Table 1).

The quantum chemist or the computational molecular physicist could then be periodically called upon to provide the specified residues' structural, energetic, orbital and electron density results from his or her model peptides.

At the *present* time, we chemical, molecular, quantum and theoretical physicists should be wary of any one model attempting to fully describe the folding of polypeptides into biologically functional proteins. All the while, approaching even our own works with skepticism, demanding reproducibility and non *statistically-massaged* proof.

The questions and challenges inherent in this problem require the combined interpretation of results emerging from all disciplines. A standardised definition and computed peptide structure database [93,119] must be established, able to be used by all investigators from medicine and biology, through chemistry and physics to computer science and informatics.

A race towards a full understanding, using solely one approach will otherwise result in a large amount of repetition of works, due to a lack of communication between differing disciplines.

The lure of simple models promising to incorporate all elements and driving forces responsible for the folding of polypeptides into biologically functional proteins, must be avoided and the results interpreted with caution. Simple models will certainly provide answers to a large number of questions, however, they may also be devoid of finely tuned and highly specific directing forces, with the models effectively *skewing the fold* to a structure proximate in conformational hyperspace to the native one, but with a reduction-in or absence-of biological activity.

> We should be careful to get out of an experience only the wisdom that is in it — and stop there; lest we be like the cat that sits down on a hot stove-lid. She will never sit down on a hot stove-lid again; but also she will never sit down on a cold one any more.
>
> – Mark Twain –

A repeated, iterative refinement of the simple models must therefore be undertaken, until the process suffers from the law of diminishing returns whereby the difference between the results of successively refined models, becomes negligible.

The simple model that allows for *future* inclusion of more refined reductionist terms, even if only in practice, will be afforded with a folding path to the true bases of native structure and activity.

EPILOGUE

We finish this tribute to Alberto Galindo quoting what the hungarian Nobel Laureate Albert Szent-Györgyi said half a century ago and published in 1960 [120,121]:

> *The distance between those abstruse quantum mechanical calculations and the patient bed may not be as great as believed.*

Aknowledgements: We would like to thank M. Bueno, P. Bruscolini, A. Cruz, E. Freire, V. Laliena, J. Onuchic, M. Orozco, A. Rey, J. Sancho, D. H. Setiadi and our colleagues of the chemistry section in the Zaragoza University for illuminating discussions. This work has been partially supported through the following research contracts: grant BFM2003-08532-C02-01, MCYT (Spain) grant FPA2001-1813, Grupo Consolidado Aragón Government "Biocomputación y Física de Sistemas Complejos" and the Global Institute Of COmputational Molecular and Materials Science (GIOCOMMS, Toronto, Canada). P. Echenique is supported by MECD (Spain) FPU grant. I. G. Csizmadia wishes to thank the Hungarian Ministry of Education for a Szent-Györgyi Visiting Professorship.

REFERENCES

1. Terradas E. and Ortiz R., *Relatividad*, Argentina: Espasa-Calpe S. A., 1952.
2. Galindo A., *Nuovo Cimento* **37**, 413 (1965).
3. Galindo A., *Actas de la primera reunión de Física Teórica*, Santander, 1965, vol. 2.
4. Anfinsen C. B., *Science* **181**, 223 (1973).
5. Ramsen I., *Unsolved problems in chemistry*, in *Modern inventions and discoveries*, New York: J. A. Hill and Co., 1904. Also quoted in Neurath N., *Protein Science* **1**, 2 (1992).
6. Quote from Oriol Bohigas at the 1st BIFI meeting, *La complejidad como paradigma: de la Física a la Biología*, 6th April 2001, Instituto de Ciencias de Materiales, CSIC, Universidad Autónoma de Madrid.
7. Laughlin R. B., Pines D., Schmalian J., Stojkovic B. P. and Wolynes P., *Proc. Natl. Acad. Sci.* **97**, 32 (2000).
8. Schrödinger E., *What is life? The Physical Aspect of the Living Cell*, Cambridge. Mass.: Cambridge University Press, 1944.

9. Dill K. A., *Protein Sci.* **8**, 1166 (1999).
10. Dawkins R., *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*, USA Publisher: W. W. Norton and Company, 1986.
11. Chan H. S., Dill K. A., *J. Chem. Phys.* **95**, 3775 (1991).
12. Dill K. A., Bromberg S., Yue K., Fiebig K. M., Yee D. P., Thomas P. D. and Chan H. S., *Protein Sci.* **4**, 561 (1995).
13. Matthews B. W., *Annu. Rev. Biochem.* **62**, 139 (1993).
14. Levinthal C., *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House*, Monticello, Illinois, Editors: DeBrunner J. T. P., Munck E., University of Illinois Press, 1969, pp. 22-24. (see also `http://brian.ch.cam.ac.uk/~mark/levinthal/levinthal.html`)
15. Hardin C., Eastwood M. P., Prentiss M., Luthey-Shulten Z. and Wolynes P., *J. Comput. Chem.* **23**, 138 (2002).
16. Bryngelson J. D. and Wolynes P. G., *Proc. Natl. Acad. Sci.* **84**, 7574 (1981).
17. Bryngelson J. D. and Wolynes P. G., *J. Phys. Chem.* **93**, 6902 (1989).
18. Cheung M. S., García A. and Onuchic J. N., *Proc. Natl. Acad. Sci.* **99**, 685 (2002) and references therein.
19. Wolynes P. G., Onuchic J. N. and Thirumalai D., *Science* **267**, 1619 (1995).
20. Anfinsen G. B. and Scheraga H. A., *Adv. Prot. Chem.* **29**, 209 (1975).
21. Kim P. S. and Baldwin R. L. *Annu. Rev. Biochem.* **59**, 631 (1990).
22. Matthews C. R., *Annu. Rev. Biochem.* **62**, 653 (1993).
23. Dobson C. M., Evans P. A. and Radford S. E., *Trends Biochem Sci* **19**, 31 (1994).
24. Kuwajima K., *Proteins* **6**, 87 (1989).
25. Ptitsyn O. B., *FEBS Lett* **285**, 176 (1991).
26. Csermely P., *BioEssays* **21**, 959 (1999).
27. Ptitsyn O. B., *Adv. Prot. Chem.* **47**, 83 (1995).
28. Levitt M., Gerstein M., Huang E., Subbiah S. and Tsai J., *Annu. Rev. Biochem.* **66**, 549 (1997).
29. Raina S. and Missiakas D., *Annu. Rev. Microbiol.* **51**, 179 (1997).
30. Hamilton G. S. and Steiner J. P., *J. Med. Chem.* **41**, 5119 (1998).
31. Székely Z.,Konya Z., Becskei A., Goldring W. P. D., Perczel A, Penke B., Molnár J., Michejda C. J., Aszalós A. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **367**, 159 (1996).
32. Viviani W., Loos M., Rivail J.-L., Csaszar P., Perczel A. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **455**, 107 (1998).
33. Baldoni H. A., Zamarbide G. N., Enriz R. D., Jáuregui E. A., Farkas O., Perczel A., Salpietro S. J. and Csizmadia I. G., *J. Mol. Struct (Millennium Volume)* **500**, 97 (2000).
34. Sahai M. A., Kehoe T. A. K., Koo J. C. P., Setiadi D. H., Chass G. A., Penke B., Pai E. F. and Csizmadia I. G., *To be published*
35. Abagyan R. A. and Totrov M., *J. Comput. Phys.* **151**, 402 (1999).
36. Ramachandran G. N. and Ramakrishnan C., *J. Mol. Biol.* **7**, 95 (1963).
37. Zimm B. H. and Bragg J. K., *J. Chem. Phys.* **31**, 526 (1959).
38. Lifson S., Roig A., *J. Chem. Phys.* **34**, 1963 (1961).
39. Doig A. J., *Biophys. Chem.* **101-102**, 281 (2002).

40. Muñoz V. and Serrano L., *Nat. Struct. Biol.* **1**, 399 (1994).
41. Muñoz V. and Serrano L., *Biopolymers* **41**, 495 (1997).
42. Flory P. J., *Statistical Mechanics of Chain Molecules*, New York: Wiley, 1969, p. 252.
43. Ramachandran G. N. and Ramakrishnan C., *J. Mol. Biol.* **7**, 95 (1963).
44. Chan H. S., Bromberg S. and Dill K. A., *Philos. Trans. R. S. London* **348**, 61 (1995).
45. Pappu R. V., Srinivasan R. and Rose G. D., *Proc. Natl. Acad. Sci.* **97**, 12565 (2000).
46. Srinivasan R. and Rose G. D., *Proc. Natl. Acad. Sci.* **96**, 14258 (1999).
47. Yu C. H., Norman M. A., Schäfer L., Ramek M., Peeters A. and Van Alsenoy C., *J. Mol. Struct.* **567-568**, 361 (2001).
48. Perczel A., Farkas O., Jakli I., Topol I. A. and Csizmadia I. G., *J. Comp. Chem.* **24**, 1026 (2003).
49. Vargas R., Garza J., Hay B. P. and Dixon D. A., *J. Phys. Chem. A* **106**, 3213 (2002).
50. Iwaoka M., Okada M. and Tomoda S., *J. Mol. Struct. (THEOCHEM)* **586**, 111 (2002).
51. Topol I. A., Burt S. K., Deretey E., Tang T. H., Perczel A., Rashin A. and Csizmadia I. G., *J. Am. Chem. Soc.* **123**, 6054 (2001).
52. Pauling L., Corey R. B. and Branson H. R., *Proc. Natl. Acad. Sci.* **37**, 235 (1951).
53. Ben-Tal N., Sitkoff D., Topol I. A., Yang A.-S., Burt S. K. and Honig B., *J. Phys. Chem. B* **101**, 450 (1997).
54. Galzitskaya O. V. and Finkelstein A. V., *Proc. Natl. Acad. Sci.* **96**, 11299 (1999).
55. Muñoz V. and Eaton W. A., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11311 (1999).
56. Bruscolini P. and Pelizzola A., *Phys. Rev. Lett.* **88**, 258101 (2002).
57. Bruscolini P. and Cecconi F., *J. Chem. Phys.* **119**, 1248 (2003).
58. Onuchic J. N., Wolynes P. G., Luthey-Schulten Z. and Socci N. D., *Proc. Natl. Acad. Sci.* **92**, 3626 (1995).
59. Hoang T. X., Trovato A., Seno F., Banavar J. R. and Maritan A., *Proc. Natl. Acad. Sci.* **101**, 7960 (2004).
60. Dill K. A. and Bromberg S., *Molecular Driving Forces*, New York and London: Garland Science, Taylor and Francis Group, 2003.
61. Ramachandran G. N. and Sasiekharan V., *Adv. Prot. Chem.* **23**, 283 (1968).
62. Baldwin R. L. and Rose G. D., *Trends Biochem. Sci.* **24**, 26 (1999).
63. Eisenberg D., *Proc. Natl. Acad. Sci.* **100**, 11207 (2003).
64. Kauzmann W., *Adv. Prot. Chem.* **14**, 1 (1959).
65. Abagyan R. A. in *Computer Simulations of Biomolecular Systems*, edited by Van Gusteren W. F., Kluwer academic publishing, 1997, vol. 3.
66. Van Gusteren W. F. and Berendsen H. J. C., *Mol. Phys.* **34**, 1311 (1977).
67. Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S. and Karplus M., *J. Comput. Chem.* **4**, 187 (1983)
68. Mazur A. K., Doroteev V. E. and Abagyan R. A., *J. Comput. Phys.* **92**, 261 (1991).
69. Levitt M. and Warshell A., *Nature* **253**, 694 (1975).
70. Go N. and Taketomi H., *Proc. Natl. Acad. Sci.* **75**, 559 (1978).
71. Miyazawa S. and Jernigan R. L., *Macromulecules* **18**, 534 (1985).

72. Srinivasan R. and Rose G. D., *Proteins* **22**, 81 (1995).
73. Kolinski A. and Skolnick J., *Proteins* **18**, 338 (1994).
74. Rykunov D. S., Reva B. A. and Finkelstein A. V., *Proteins* **22**, 100 (1995).
75. Abagyan R. A. and Totrov M., *J. Mol. Biol.* **235**, 983 (1994).
76. Pedersen J. T. and Moult J., *Proteins* **23**, 454 (1995).
77. Auspurger J. D. and Scherga H. A., *J. Comput. Chem.* **17**, 1549 (1996).
78. Alonso J. L., Echenique P. and Tarancón A., *in preparation.*
79. Perczel A., Farkas O. and Csizmadia I. G., *J. Am. Chem. Soc.* **117**, 1653 (1995).
80. Tanford C., *Proc. Natl. Acad. Sci.* **76**, 4175 (1975).
81. Sharp K., Nicholls A., Friedman R. and Honig B., *Biochemistry* **30**, 9696 (1991).
82. Chan H. S., *Amino acid side-chain hydrophobicity* in *Encyclopedia of Life Sciences*, Macmillan Publishers Ltd., Nature Publishing Group, 2002.
83. Eisenberg D. and McLachlan A. D., *Nature* **319**, 1999 (1986).
84. Lee B. and Richards F. M., *J. Mol. Biol.* **55**, 379 (1971).
85. Zagrovic B., Sorin E. J. and Pande V., *J. Mol. Biol.* **313**, 151 (2001).
86. Jorgensen W. L. and Tirado-Rives J., *J. Am. Chem. Soc.* **110**, 1666 (1988).
87. Qiu D., Shenkin P. S., Hollinger F. P. and Still W. C., *J. Phys. Chem.* **101**, 3005 (1997).
88. Muñoz V., Thompson P. A., Hofrichter J. and Eaton W. A., *Nature* **390**, 196 (1997).
89. Muñoz V., Henry E. R., Hofrichter J. and Eaton W. A., *Proc. Natl. Acad. Sci.* **95**, 5872 (1998).
90. Chass G. A., Rodriguez A. M., Mak M. L., Deretey E., Perczel A., Sosa C. P., Enriz R. D. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **537**, 319 (2001).
91. Perczel A., Angyan J. G., Kajtar M., Viviani W., Rivail J.-L., Marcoccia J. F. and Csizmadia I. G., *J. Am. Chem. Soc.* **113**, 6256 (1991).
92. Perczel A. and Csizmadia I. G. *Ab initio conformational analysis of protein subunits* in *The Amide Linkage; Structural Significance in Chemistry, Biochemistry and Materials Science*, Edited by Greenberg A., Breneman C. M. and Liebman G. F., Wiley Interscience, 200, pp. 409-461.
93. Chass G. A., Sahai M. A., Law J. M. S., Lovas S., Farkas O., Perczel A., Rivail J.-L. and Csizmadia I. G., *Int. J. Quant. Chem.* **90**, 933 (2002).
94. Perczel A., McAllister M. A., Császsr P. and Csizmadia I. G., *J. Am. Chem. Soc.* **115**, 4849 (1993).
95. McAllister M. A., Perczel A., Császár P. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **288**, 181 (1993).
96. Perczel A., McAllister M. A., Császár P. and Csizmadia I. G., *Can. J. Chem.* **72**, 2050 (1994).
97. Borics A., Chass G. A., Csizmadia I. G., Murphy R. F. and Lovas S., *J. Mol. Struct. (THEOCHEM)* **666-667**, 355 (2003).
98. Patel M. A., Deretey E. and Csizmadia I. G., *J. Mol. Struct (THEOCHEM)* **492**, 1 (1999).
99. Cheung M., McGovern M. E., Jin T., Zhao D. C., McAllister M. A., Perczel A., Császár P. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **309**, 151 (1994).
100. Sahai M. A., Sahai M. R., Chass G. A., Penke B. and Csizmadia I. G., *J. Mol. Struct. (THEOCHEM)* **666-667**, 327 (2003).

101. Chass G. A., Marai C. N. J., Setiadi D. H., Harrison A. G. and Csizmadia I. G. *J. Mol. Struct. (THEOCHEM)* **675**, 149 (2004).
102. Chass G. A., Marai C. N. J., Harrison A. G. and Csizmadia I. G., *J. Phys. Chem. A* **106**, 9695 (2002).
103. Chin W., Mons M., Dognon J.-P., Piuzzi F., Tardivel B. and Dimicoli I., *Phys. Chem. Chem. Phys.* **6** (2004).
104. Mirasol R., Setiadi D. H., Tang T.-H., Chin W., Mons M., Dimicoli I., Dognon J.-P., Viskolcz B., Lovas S., Penke B., Chass G. A. and Csizmadia I. G., *J. Phys. Chem. A*, submitted.
105. Chin W., Mons M., Dognon J.-P., Mirasol R., Chass G. A., Dimicoli I., Piuzzi F., Butz P., Tardivel B., Compagnon I., von Helden G. and Meijer G., *J. Phys. Chem. A*, submitted.
106. Burkley S. K. and Petsko G. A., *Adv. Prot. Chem.* **39**, 125 (1988).
107. Chipot C., Maigret B., Pearlman D. A. and Kollman P. A., *J. Am. Chem. Soc.* **118**, 2998 (1996).
108. McGaughey G. B., Gagnes M. and Rappe A. K., *Proteins* **273**, 15458 (1998).
109. Tsuzuki S., Honda K., Uchimaru T., Mikami M. and Tanabe K., *J. Am. Chem. Soc.* **122**, 3746 (2000).
110. Toth G., Murphy R. F. and Lovas S., *J. Am. Chem. Soc.* **123**, 11782 (2001).
111. Toth G., Murphy R. F. and Lovas S., *Prot. Engineer.* **14**, 543 (2001).
112. Steiner T. and Koellner G., *J. Mol. Biol.* **305**, 535 (2001).
113. Shi Z., Olson S. A. and Kallenbach N. R., *J. Am. Chem. Soc.* **124**, 3284 (2002).
114. Chass G. A., Lovas S., Murphy R. F. and Csizmadia I. G., *Eur. Phys. J. D* **20**, 481 (2002).
115. Chass G. A., Mirasol R. S., Setiadi D. H., Marai C. N. J., Tang T. H., Murphy R. F., Csizmadia I. G. and Lovas S., *J. Mol. Struct. (THEOCHEM)*, submitted.
116. Eberhardt E. S., Panisik N. J. and Raines R. T., *J. Am. Chem. Soc.* **118**, 12261 (1996).
117. Milner-White E. J., Bell L. H. and Maccallum P. H., *J. Mol. Biol.* **228**, 725 (1992).
118. Vitagliano L., Berisio R., Mastrangelo A., Mazzarella L. and Zagari A., *Prot. Sci.* **10**, 2627 (2001).
119. Sahai M. A., Lovas S., Chass G. A., Penke B. and Csizmadia I. G. *J. Mol. Struct. (THEOCHEM)* **666-667**, 169 (2003).
120. Szent-Györgyi A., *Introduction to Submolecular Biology*, New York and London: Academic Press, 1960, p. 114.
121. Csizmadia I. G., Penke B. and Toth G. (Ed.), *The Role of Chemistry in the Evolution of Molecular Medicine. A Tribute to Professor Albert Szent-Györgyi*, Amsterdam: Elsevier, 2004, p. 856. (This publication includes a reproduction of *Introduction to Submolecular Biology* by Albert Szent-Györgyi)